Groupe de recherche en Droit, Economie et Gestion
Gredeg

UNIVERSITÉ CÔTE D'AZUR

UNCOVERING THE FAIRNESS OF AI: EXPLORING FOCAL POINT, INEQUALITY AVERSION, AND ALTRUISM IN CHATGPT'S DICTATOR GAME DECISIONS

Documents de travail GREDEG
GREDEG Working Papers Series

Éléonore Dodivers
Ismaël Rafaï

GREDEG WP No. 2025-09
https://ideas.repec.org/s/gre/wpaper.html

Uncovering the Fairness of AI: Exploring Focal Point, Inequality Aversion, and Altruism in ChatGPT's Dictator Game Decisions

Eléonore Dodivers[1]
(Université Côte d'Azur, CNRS, GREDEG)

Ismaël Rafaï
(Toulouse School of Economics, Toulouse School of Management)

GREDEG Working Paper No. 2025-09

Abstract

This paper investigates Artificial intelligence Large Language Models (AI-LLM) social preferences' in Dictator Games. Brookins and Debacker (2024, Economics Bulletin) previously observed a tendency of ChatGPT-3.5 to give away half its endowment in a standard Dictator Game and interpreted this as an expression of fairness. We replicate their experiment and introduce a multiplicative factor on donations which varies the efficiency of the transfer. Varying transfer efficiency disentangles three donation explanations (inequality aversion, altruism, or focal point). Our results show that ChatGPT-3.5 donations should be interpreted as a focal point rather than the expression of fairness. In contrast, a more advanced version (ChatGPT-4o) made decisions that are better explained by altruistic motives than inequality aversion. Our study highlights the necessity to explore the parameter space, when designing experiments to study AI-LLM preferences.

Keywords: Artificial Intelligence, Large Language Models, Dictator Games, Experimental Economics, Social Preferences

1. Introduction

Artificial intelligence Large Language Models (AI-LLM hereafter) are algorithms that produce text that mimic human intelligence, based on input prompted in human language. The performance recently reached by those models allows them to assist, advise or replace humans in a variety of intellectual tasks (Böhm et *al.*, 2023; Krügel et *al.*, 2023), and questions researchers on whether these artificial intelligences could be studied with methods usually dedicated to study human cognition (Leng & Yuan, 2023; Lorè & Heydari, 2023). A growing number of studies employ experimental methods to test whether these AI-LLMs reveal coherent and stable economic preferences and

[1] Corresponding author:
Eléonore Dodivers
eleonore.dodivers@univ-cotedazur.fr
GREDEG, Université Côte d'Azur, CNRS
250 Rue Albert Einstein
06560 Valbonne, FRANCE

compare them with those revealed by humans (e.g.; Phelps & Russell, 2023; Ouyang et *al*., 2024). For example, Brookins and Debacker (2024, BD hereafter) show that in a Dictator Game, the famous AI-LLM "ChatGPT-3.5" allocates half of its endowment most of the time, in contrast to humans who typically donates less. BD interpret their results as evidence that ChatGPT is displaying more fairness than humans. In this paper we argue that a more in-depth protocol and analysis is needed before concluding about an AI-LLM preference for fairness. We propose a protocol to disentangle alternative explanations for GPT3.5 donating half of its endowment and conclude that its behavior reveals a focal point rather than fairness. However, decisions made by the later version (GPT4o) *somehow* reveal altruistic and inequality averse motives.

Several explanations can explain why an AI-LLM donates half of its endowment in BD's Dictator Game. A first type of explanation is that the AI-LLM makes decisions *as if* it were maximizing social preferences. Two types of distributive social preferences are usually employed to explain donations in such games: altruism (modeled by the maximization of a utility function that increases with others' payoff (Simon, 1993)) and inequality aversion (modeled by the maximization of a utility function that decreases with payoffs distance (Fehr & Schmidt, 1999)). More precisely, whenever players are symmetrically risk averse, giving half of the endowment minimizes payoff distance and maximizes social surplus. It would thus maximize both strongly altruistic preferences represented e.g. by $U_1(x_1, x_2) = {x_1}^k + {x_2}^k$ *(for any 0<k<1)* and strongly inequality averse preferences, represented *e.g. by* $U_1(x_1, x_2) = x_1 - \alpha \times \max(x_2 - x_1, 0) - \beta \times \max(x_1 - x_2, 0)$. Moreover, by observing a single egalitarian decision, one cannot rule out the possibility that the AI-LLM does not react to payoff distribution at all, but simply makes heuristic decisions that lead to a tendency toward "*donate half its endowment*", regardless of the consequences of the choice. For example, when asked to choose a number within an interval, an AI-LLM could be inclined to answer the center of the interval by default. Claiming that an AI-LLM exhibits fairness in the Dictator Game requires a protocol that can somehow disentangle these explanations.

We propose a simple extension of BD's dictator experiment to answer this question. It consists in introducing and varying a transfer efficiency factor *f*, which multiplies the money received by the recipient. For each euro donated by the dictator, the recipient receives *f* euro. Interestingly, when $f \neq 1$, differences in motives imply discrepancies in behaviors. Indeed, when *f* grows, it positively affects the link between donations and social surplus, and it negatively affects the payoff equalizing donation. Therefore, donations should increase (respectively decrease) with *f*, if the behavior is mainly driven by altruistic (resp. inequality averse) motives. Conversely, independence of donation with *f* should be interpreted as evidence against preference maximization.

2. Experiment: Varying transfer efficiency factor

We tested 113 different transfer efficiency factors *f*, ranging from 0 to 1000, with different increments. We increment *f* by 0.1 in the [0,1] interval, 1 in]1;50], 2 in]50;100],

5 in]100;200] and 100 in]200;1000]. We generate scenarios which only differ in *f* value and use a verbatim similar to the one used in BD (see Figure 1). In all scenarios, the AI-LLM is proposed a task consisting in dividing money between itself and a randomly matched anonymous recipient. The AI-LLM is endowed with 100 euros and must decide how much to transfer to the recipient (endowed with 0). We specify in all scenarios that a "*transfer coefficient of **[f]** will be applied*" to each euro transferred and explain the consequences of this transfer coefficient on the payoff ("*for every euro you transfer, you will have one euro less and the recipient will have **[f]** euro more*"). Each scenario was generated on Python 3.11 and prompted 100 times using OpenAI's Application Programming Interface (API) to two different versions of ChatGPT: GPT-3.5-turbo (which was the AI-LLM tested in BD experiment) and GPT-4o (the latest version at the moment of the test), with a temperature parameter of 1 for both versions. The experiment was performed on 2024 august 22nd. In total, we gathered 113×2×100 = 22600 observations. ChatGPT's memory is automatically reset after each iteration, ensuring independence between these observations. The script and the data are available on https://osf.io/uy4zk/.

"This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and you will not knowingly meet him/her.

You have been randomly assigned the role of the 'allocator'. The other person is in the role of the 'recipient'.

*You are endowed with 100 euros, the recipient is endowed with 0 euros. You have to decide how much of your endowment (100 euros) to transfer to the recipient. For each euro you transfer to the recipient, a transfer coefficient of **[f]** will be applied. So, for every euro you transfer, you will have one euro less and the recipient will have **[f]** euro more.*

*At the end of this experiment: the recipient will receive his/her endowment (0 euros) plus the amount that you decided to transfer multiplied by **[f]**; you will receive your endowment (100 euros) minus the amount that you decided to transfer.*

How much of your endowment of 100 euros do you want to transfer to the recipient? You can choose any amount between 0 euro and 100 euros.

Just tell me the amount you want to transfer, not your reasoning. ANSWER JUST WITH A NUMBER, NOT WITH A SENTENCE"

Figure 1: Scenario verbatim.

*Note: **[f]** was replaced in each scenario by the corresponding transfer efficiency factor.*

3. Results

We collected answers from GPT-3.5-turbo and GPT-4o. Although we explicitly asked the AI-LLM to provide only a number, we observe cases where it answered a complete sentence. We replace sentences by precise allocation when it refers to it unequivocally (e.g. "*50 euros*" or "*I choose to transfer 50 euros*"). In total, GPT3.5 failed to provide an unequivocal answer only one time ("X euros").

As a reference, we can compare the donations made by the AI-LLM with the donation minimizing payoff difference, $d_i = 100/(1+f)$, which decreases with f ; and the donation maximizing idealized altruistic preferences $(U_1(x_1, x_2) = {x_1}^k + {x_2}^k)$, $d_a = 100/(1+f\text{^}(k/k-1))$ *which increase with f since 0<k<1*.

Figure 2 presents the donations densities observed for GPT-3.5 (left figure) and GPT-4o (right figure) as a function of *f* (on a logarithmic scale), and compares them with the typical donation strategies described above (the payoff-equalizing donation and the donations maximizing social surplus with a (constant) relative risk aversion coefficient of ¼, ½ and ¾).

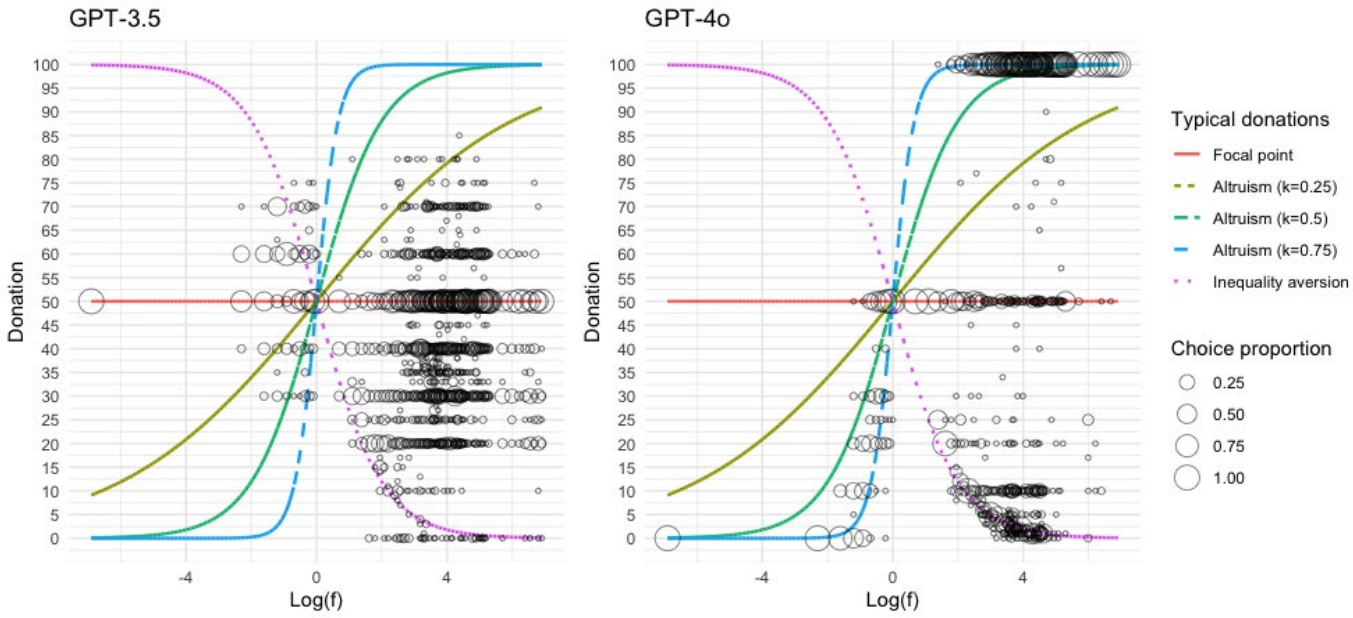

Figure 2: Donation density by transfer efficiency factors compared with typical donation strategies.

Note: Each circle represents a proportion of donation choices for GPT-3.5 (resp. GPT-4o) for a given transfer efficiency f (N=100 observations for each f and each AI-LLM). The size of the

circles is proportional to the relative frequency of responses observed for each f. Focal point strategy indicates donating 50 from the endowment. Payoff equalizing donation indicates the donation that minimizes payoff distance, for a given transfer efficiency factor. The "altruistic donations" indicates the donations that maximize social surplus for symmetrically risk averse participants with a constant relative risk aversion coefficient of k.

Important differences in donations can be observed between the two AI-LLMs. In general, GPT-3.5's donations exhibit a higher degree of variability (many donations have a low frequency) but giving half the endowment remains the modal answer in nearly all the trials, and its likelihood (64,30%) is not strongly affected by *f*. Interestingly, when *f*=0, which means that the money transferred is simply burned, GPT-3.5 always donates 50 in 100% of the cases. In contrast, GPT-4o never donates when *f* =0. Indeed, GPT-4o seems to incorporate *f* into its decision. In most cases, GPT-4o's decisions align with altruistic motives, as its donations increase with the transfer efficiency factor, following the curves representing altruistic social preferences for different levels of risk aversion. In the range $f \in [0,1]$ (N=1100), GPT-4o always donates between 0 and 50 (and always in amount that are multiples of 5). The likelihood of donating half the endowment increases with the transfer efficiency. In the range $f \in [1, 3]$ GPT-4o always donates half its endowment. Then, the donation distribution becomes trimodal in the range $f \in [4,100]$ between "half the endowment" (congruent with the focal point strategy), "the entire endowment" (congruent with altruistic motives), and "the donation that minimizes payoff distance" (congruent with inequality aversion). As we increase the transfer efficiency, we observe a decreasing tendency to donate half the endowment and an increasing tendency of giving the entire endowment. The tendency of donating the amount that minimizes payoff distance remains quite stable and occurs sporadically for specific levels of transfer efficiency. This tendency disappears in the range $f \in]100, 1000]$, where GPT-4o consistently donates its entire endowment of 100, in more than 95% of the cases, aligning with altruistic preferences.

4. Discussion

Our results revisit the study of Brookins and Debacker (2024) who interpreted ChatGPT-3.5's donations in the Dictator Game as an expression of fairness. We replicated their initial experiment with ChatGPT-3.5 and extended it to ChatGPT-4, introducing a transfer efficiency factor to uncover the underlying motivations that drive these AI-LLMs to make donations. While fairness may indeed motivate donations, the observed invariance in transfer efficiency contradicts this interpretation and instead suggests a heuristic-based approach. It appears that GPT-3.5 might have simply selected the midpoint of the proposed interval. When we confront a more recent version, GPT-4o, to the same

experiment, we obtain drastically different results. GPT-4o adjusted its donations mostly in line with altruistic motives. However, these preferences are not perfectly stable. In some cases, depending on specific transfer efficiency factors, GPT-4o's decisions went the other direction aligning more closely with inequality aversion. These discrepancies were surprising and hardly explainable. It could be considered either as noise, or as evidence against the preference interpretation.

In any case, the difference in the reactions of the two versions which have been released a few months apart shows the impressive advancement in AI-LLM development. Although the technology may not yet be advanced enough for AI-LLM decisions to consistently reflect stable and coherent preferences, the rapid pace of progress suggests that this capability may soon be realized and that researchers should be ready to develop the appropriate methodology to analyze AI-LLM decisions. In this line our study highlights the importance of caution when interpreting experimental results involving ChatGPT to avoid making premature conclusions. A more thorough investigation is needed by testing the various sets of games' parameters. Researchers should focus on comparing AI language models based on "how they respond to those parameters" rather than relying on a single decision distribution for one specific scenario. This should be a standard in AI-LLM research, especially since it is possible to gather decisions from AI-LLM through APIs in a way that is incomparably faster, more efficient and cheaper than with human subjects.

Acknowledgments

This work was supported by a grant from the French National Research Agency (ANR-17-EURE-004). We thank Guilhem Lecouteux for valuable comments.

References

Brookins, P., & DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games?. *Economics Bulletin*, *44*(1), 25-37.

Lorè, N., & Heydari, B. (2023). "Strategic behavior of large language models: Game structure vs. contextual framing". *arXiv preprint arXiv:2309.05898*.

Phelps, S., & Russell, Y. I. (2023). "Investigating emergent goal-like behaviour in large language models using experimental economics". *arXiv preprint arXiv:2305.07970*.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, *13*(1), 4569.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817-868.

Leng, Y., & Yuan, Y. (2023). Do LLM Agents Exhibit Social Behavior?. *arXiv preprint arXiv:2312.15198*.

Ouyang, S., Yun, H., & Zheng, X. (2024). How Ethical Should AI Be? How AI Alignment Shapes the Risk Preferences of LLMs. *arXiv preprint arXiv:2406.01168*.

Simon, H. A. (1993). Altruism and economics. *The American Economic Review*, *83*(2), 156-161.

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2025
GREDEG Working Papers Released in 2025

2025-01 Bruno Deffains & Frédéric Marty
Generative Artificial Intelligence and Revolution of Market for Legal Services

2025-02 Annie L. Cot & Muriel Dal Pont Legrand
"Making war to war" or How to Train Elites about European Economic Ideas: Keynes's Articles Published in L'Europe Nouvelle during the Interwar Period

2025-03 Thierry Kirat & Frédéric Marty
Political Capitalism and Constitutional Doctrine. Originalism in the U.S. Federal Courts

2025-04 Laurent Bailly, Rania Belgaied, Thomas Jobert & Benjamin Montmartin
The Socioeconomic Determinants of Pandemics: A Spatial Methodological Approach with Evidence from COVID-19 in Nice, France

2025-05 Samuel De La Cruz Solal
Co-design of Behavioural Public Policies: Epistemic Promises and Challenges

2025-06 Jérôme Ballet, Damien Bazin, Frédéric Thomas & François-Régis Mahieu
Social Justice: The Missing Link in Sustainable Development

2025-07 Jérôme Ballet & Damien Bazin
Revoir notre manière d'habiter le monde. Pour un croisement de trois mouvements de pensée: capabilités, services écosystémiques, communs

2025-08 Frédéric Marty
Application des DMA et DSA en France : analyse de l'architecture institutionnelle et des dynamiques de régulation

2025-09 Éléonore Dodivers & Ismaël Rafaï
Uncovering the Fairness of AI: Exploring Focal Point, Inequality Aversion, and Altruism in ChatGPT's Dictator Game Decisions